【ICDE2016 & WWW2016 勉強会】

ICDE 2016 Session 2B: Beyond Relational Query Processing

Hobbes3: Dynamic Generation of Variable-Length Signatures for Efficient Approximate Subsequence Mappings

清水敏之（京都大学）

※ 図表は論文より引用

Hobbes3: Dynamic Generation of Variable-Length Signatures for Efficient Approximate Subsequence Mappings

Jongik Kim (Chonbuk National University), Chen Li (University of California), Xiaohui Xie (University of California)

- やりたいこと: DNA配列の高速部分マッチング(部分文字列マッチング)
  - read mapping
- アプローチ: signatures(クエリ文字列の部分文字列集合)を用いて候補を絞り込み(filtering)→正確に一致するか確認(verification)
- 新規性: 固定長ではなく可変長のsignaturesを用いる
  - filtering powerとsignatureを作るコストのバランスを考慮して可変長signaturesを選択するアルゴリズムを提案

近似マッチングも考慮する

大枠の説明では距離関数としてhamming distanceを用いて議論されています

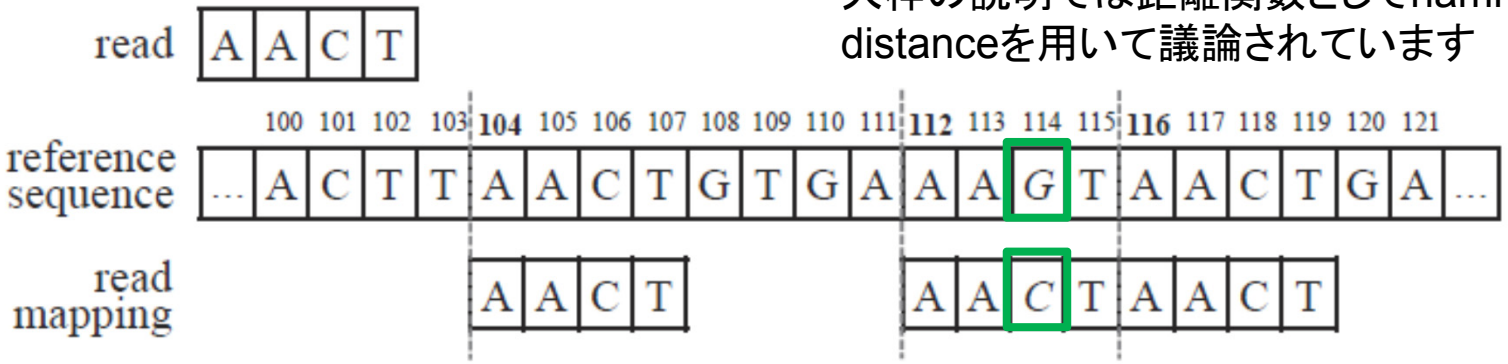


Fig. 1. An example of a read mapping

既存手法: 固定長signatures, *q*-gram

- 対象sequenceの*q*-gramを生成
  - 例: GAATGAAT の　3-gram
    → {(GAA, 0), (AAT, 1), (ATG, 2), (TGA, 3), (GAA, 4), (AAT, 5)}
- 各gramのinverted listを用意しておく

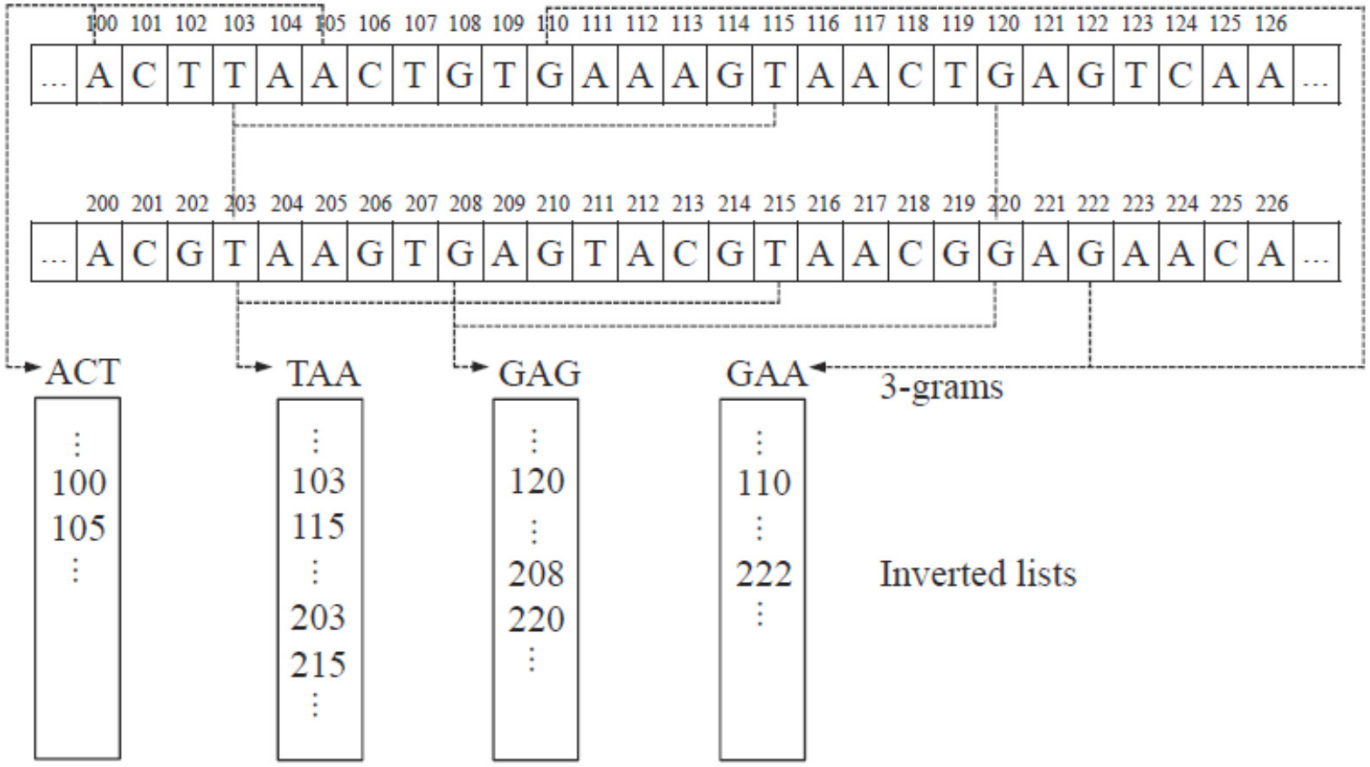
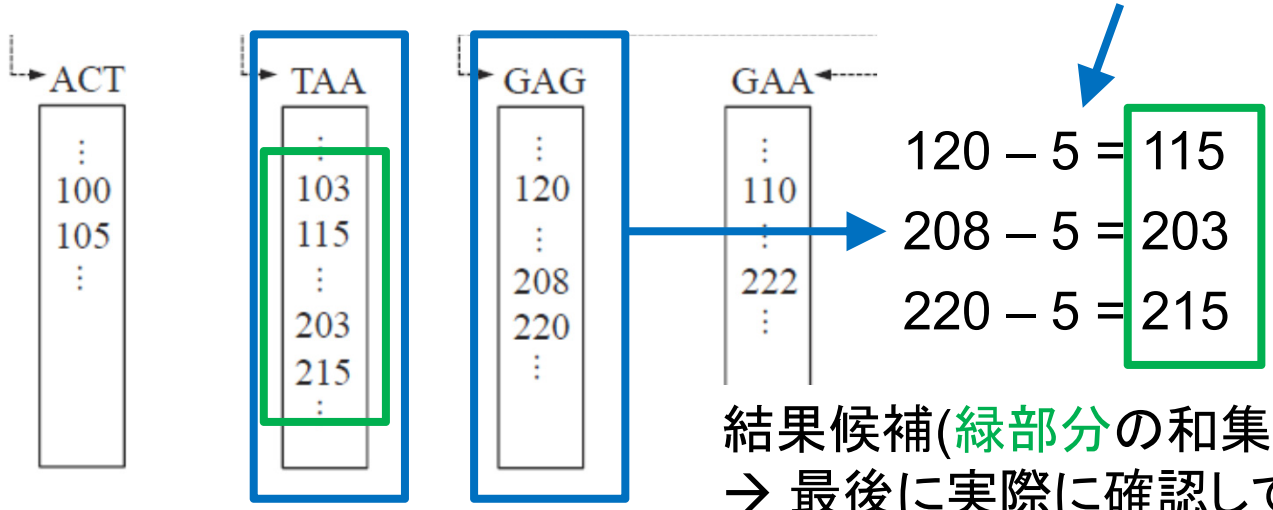


Fig. 2. Excerpt of a reference and a portion of its 3-gram inverted index

既存手法: 固定長signatures, *q*-gram

- pigeonhole principle を利用
  - *k*個までのエラーを許容する場合、*k+1*個のnon-overlapping *q*-grams をクエリから抽出
  - 各gramのinverted listの(normalizeされた)和集合が結果候補となる (いずれかのgramは必ずマッチしないといけない)

- 例) クエリ: TAACTGAGAAATTA, 3-gram, 2個までエラー許容
  - non-overlapping q-gramとして(TAA, 0), (GAG, 5), (TTA, 11) を選択



この例では対象にTTAは出現しない

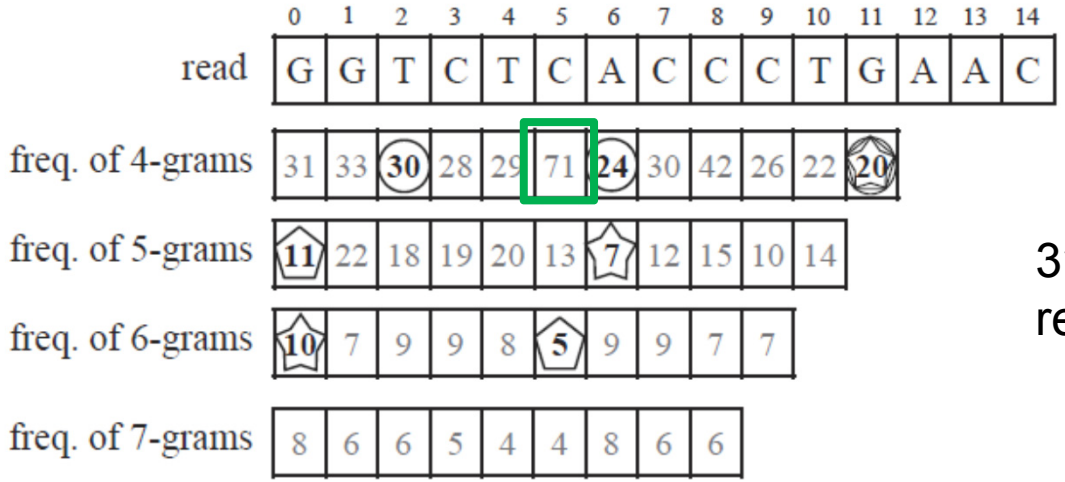
結果候補(緑部分の和集合): {103, 115, 203, 215}
→ 最後に実際に確認して103 が結果と分かる

提案手法: 可変長signatures, *q*-gram

- 大枠は固定長と同じ考え方
  - 対象sequenceは*q*-gramでinverted indexを作る(この例では4-gram)
- 複数の*q*-gramを組み合わせて可変長のsignatureを用いる
  - 可変長signatureをマッチングさせるのに(固定長の)*q*-gramsを用いるため、どのようにsignatureを選んでいくかが重要

(GGTCT, 0), (CACCCT, 5), (GAAC, 11)
filtering はよいが
(CACCCT, 5)の位置を得るのにコストが高い

(GGTCTC, 0), (ACCCT, 6), (GAAC, 11)
バランスがよい



3つのnon-overlapping 4-gramをreadから選ぶ例

Fig. 3. A read and its frequencies of q-grams in a reference sequence

- edit distanceを用いた場合の議論(4節)
- 実験で既存手法とパフォーマンスを比較(5節)